5. SIGNIFICANCE AND APPLICATION

C. E. Shannon
Bell Telephone Laboratories
Murray Hill, N. J.

1. Introduction.

A general communication system is shown in Figure 3. An information source produces a message. This is encoded in a transmitter to produce a signal suitable for transmission over the channel. During transmission the signal may be perturbed by noise. The perturbed signal is decoded or demodulated at the receiver to recover, as well as possible, the original message.

The situation is roughly analogous to a transportation system for transporting physical goods from one point to another. We can imagine, for example, a lumber mill producing lumber at an average rate of R cubic feet per second and a conveyor system capable of transporting C cubic feet per second. If R is greater than C the full output of the mill cannot possibly be carried on the conveyor. On the other hand, if R is less than or equal to C it may or may not be possible, depending on whether the lumber can be efficiently packed in the available space of the conveyer. However, if we allow ourselves to saw the lumber up into suitable sizes and shapes we can always approach 100 per cent efficiency in packing. In this case we must, of course, supply a carpenter shop at the other end of the conveyor to reassemble the lumber in its original form before passing it on to the consumer.

If the analogy is sound we might hope to define two parameters R and C associated with an information source and a channel, respectively. R should measure, in some sense, how much information is produced per second by the source, and C the capacity of the channel when used in the most efficient manner for transmitting information. We would expect then that if $R > C$ the full output of the source cannot be transmitted satisfactorily. If $R \leq C$ it should be possible to transmit the output of the source by proper encoding and decoding at transmitter and receiver. It turns out that it is possible to define quantities R and C which measure these information rates and capacities and satisfy the desired relationships. We will attempt to show how this can be done without, however, giving mathematical proofs of the results.[1]

2. The Information Source.

The first problem is that of clarifying the nature of "information" and finding a measure of the rate of production for an information source.

Information involves basically the concept of "choice." An information source chooses one particular message from a set of possible messages. If there were only

[1]For mathematical details, see Shannon, C.E., "A Mathematical Theory of Communication," Bell System Technical Journal. July and October, 1948. See also Shannon, C.E., "Communication in the Presence of Noise," Proceedings of the I.R.E. (Forthcoming).

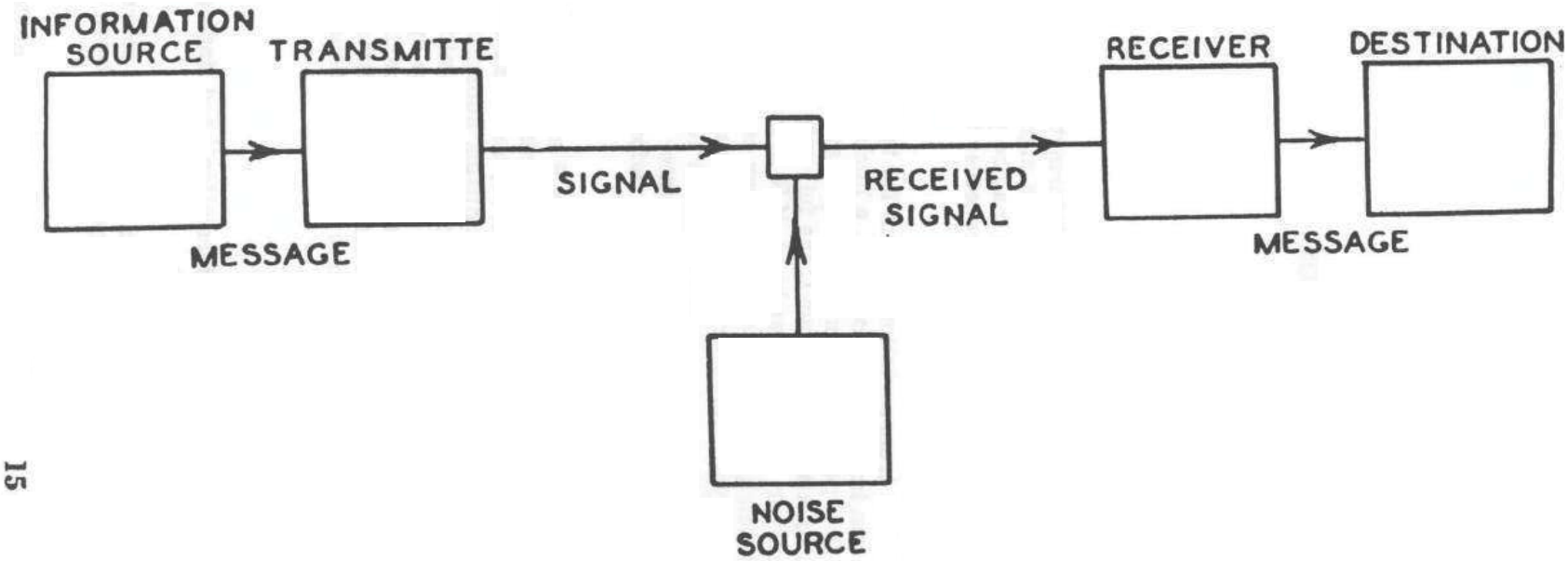
INFORMATION SOURCE
TRANSMITTE
SIGNAL
RECEIVED SIGNAL
RECEIVER
DESTINATION
MESSAGE
MESSAGE
NOISE SOURCE

Figure 3

one possible message there would be no communication problem. The amount of information produced by a source must evidently be related to the range of choice available.

The simplest possible choice is a choice from two equally likely possibilities, say 0 or 1. We shall call the corresponding unit of information a binary digit or "bit." A relay or flip-flop circuit has two possible states and is capable of storing one bit of information.

A device which chooses at random from 0 or 1 making one choice each second is considered to be producing information at rate R of one bit per second. Such a source produces a "message" which is a random sequence of 0's and 1's.

A choice from say 32 equally likely possibilities can be considered as a series of five choices, each from two equally likely possibilities, and, therefore, should correspond to five bits. More generally, a choice from n equally likely possibilities represent $\log_2 n$ bits.

Suppose now that the various possible choices have different probabilities of occurrence, say $p_1, p_2, \ldots, p_n$. How much information is produced when a choice is made under these circumstances? One feels intuitively that less "choice" is involved in a device which chooses between 0 and 1 with probabilities .01 and .99 than in one which chooses with equal probabilities. In the former case the result is almost sure to be 1.

The following example shows that by proper encoding an average compression can be obtained by using the probabilities $p_1, p_2, \ldots, p_n$. Suppose there are four possible choices A, B, C, D with probabilities $p_A = 1/2$, $p_B = 1/4$, $p_C = 1/8$, $p_D = 1/8$. If we use a simple direct code into binary digits:

$$A = 00 \qquad B = 01 \qquad C = 10 \qquad D = 11,$$

we use two binary digits per letter. On the other hand, using the following code where more probable letters are given short codes and less probable letters longer codes, we obtain an average saving

$$A = 0 \qquad B = 10 \qquad C = 110 \qquad D = 111.$$

This is a reversible code; the original text can be recovered from the encoded sequences as is readily verified. With this code we need, on the average, only

$$(1/2 \times 1 + 1/4 \times 2 + 1/8 \times 3 + 1/8 \times 3) = 1\ 3/4$$

binary digits per letter. We may say then that a choice with probabilities 1/2, 1/4, 1/8, 1/8 corresponds to 1 3/4 bits of information. If an information source were producing a sequence of the letters A, B, C, D with these probabilities we could encode it into a sequence of binary digits in which 1 3/4 binary digits are used on the average for each letter of message.

A general analysis of the situation shows that if the letters are chosen with probabilities $p_1, p_2, \ldots, p_n$ then it is possible to encode into binary digits using

$$H = - \sum p_i \log_2 p_i$$

binary digits per letter of message on the average, and there is no method of reversible encoding using less. This H then is the equivalent number of bits per letter, and, if the source produces n letters per second, R = nH is the rate of production in bits per second.

In the case of English text the statistical structure is more involved. There are the various letter probabilities p_i, but, also, there are statistical influences between nearby letters. For example, the letter T is more often followed by H than by any other letter, Q is almost invariably followed by U, etc. In such cases there is a more general formula for calculating the equivalent number of bits per letter of message. Let p(i, j, ..., s) be the probability in the language of the sequence of letters i, j, ..., s. Then we define G_n by

$$G_n = -\frac{1}{n}\sum p(i, j, \ldots, s) \log_2 p(i, j, \ldots, s)$$

where the sum is over-all sequences of letters which are just n letters long. The sequences $G_1, G_2, \ldots, G_n, \ldots$ represents a series of approximations to the desired H which takes into account more and more of the statistical structure as we proceed along the sequence. The information per letter of message can be defined by the limiting value of the G's.

$$H = \lim_{n \to \infty} G_n$$

It can be shown that H has the desired properties; namely, we can encode the messages from the source into binary digits using H binary digits per letter on the average, and no method of encoding uses less.

For the English language H has been estimated at roughly 2 bits per letter, taking account only of the statistical structure out to about 6 or 8 letters.

If the messages produced by the information source are continuous functions of time, as in speech or television transmission, the situation is much more involved and we will not discuss it in detail. It is still possible to assign a rate of production of information in bits per second to such a source, but the rate now depends on other considerations. With continuous functions as messages, exact reproduction is not generally required, and the rate R depends on the amount and nature of the discrepancy which can be tolerated between the original and recovered messages. The tolerable discrepancy in turn is determined by the final destination of the messages. With speech, for example, the tolerable errors depend on the structure of the human ear and brain.

Although the mathematical problems involved in defining the rate for a continuous source have been completely solved, it is in practical cases very difficult to estimate R. The following calculation may be of some interest, however. Suppose we are interested only in transmitting English speech (no music or other sounds), and the quality requirements on reproduction are only that it be intelligible as to meaning. Personal accents, inflections, etc., can be lost in the process of transmission. In such a case we could, at least in principle, transmit by the following scheme. A device is constructed at the transmitter which prints the English text corresponding to the spoken words. These can be translated into binary digits in the ratio of about two binary digits per letter, or 2 x 4.5 = 9 per word. Taking 100 words per minute as a reasonable talking speed we obtain 900 bits per minute or 15 bits per second as an estimate of the rate for English speech when intelligibility is the only fidelity requirement.

3. The Capacity of a Channel.

We now consider the problem of defining the capacity C of a channel for transmitting information. Since we have measured the rate of production for an information source in

bits per second, we would naturally like to measure C in the same units. The question then becomes, what is the maximum number of binary digits per second that can be transmitted over a given channel?

In some cases the answer is simple. With a teletype channel there are 32 possible symbols. Each symbol therefore represents 5 bits. If we can send n symbols per second, and the noise level is not high enough to introduce any errors during transmission, we can send 5n bits per second.

Suppose now that the channel is defined as follows. We can use for signals any functions of time f(t) which lie within a certain band of frequencies, W cycles per second wide. It is known that a function of this type can be specified by giving its value at a series of equally spaced sampling points $\frac{1}{2W}$ seconds apart. Thus we may say that such a function has 2W degrees of freedom, or dimensions, per second.

If there is no noise whatever on such a channel we can distinguish an infinite number of different amplitude levels for each sample. Consequently we could, in principle, transmit an infinite number of binary digits per second, and the capacity C would be infinite.

Even when there is noise, if we place no limitations on the transmitter power the capacity will be infinite, for we may still distinguish at each sample point an unlimited number of different amplitude levels. Only when noise is present and the transmitter power is limited in some way do we obtain a finite capacity C. The capacity depends, of course, on the statistical structure of the noise as well as the nature of the power limitation.

The simplest type of noise is white thermal noise or resistance noise. The probability distribution of amplitudes is Gaussian, and the spectrum is flat with frequency over the band in question and may be assumed to be zero outside the band. This type of noise is completely specified by giving its mean square amplitude N the power it would deliver into a unit resistance.

The simplest limitation on transmitter power is to assume that the average power delivered by the transmitter (or more precisely the mean square amplitude of the signal) be not more than P. If we define our channel by these three parameters W, P, and N, the capacity C can be calculated. It turns out to be

$$C = W \log_2 \frac{P + N}{N} \text{ (bits per second).}$$

It is easy to see that this formula is approximately right when P/N is large. The received signal will have a power P + N, and we can distinguish something of the order of

$$\sqrt{\frac{P + N}{N}}$$

different amplitudes at each sample point. In a time T there will be 2TW independent samples. Thus, there are approximately

$$M = \left(\sqrt{\frac{P + N}{N}}\right)^{2TW} = \left(\frac{P + N}{N}\right)^{TW}$$

different signal functions of duration T that can be distinguished from one another in spite of the noise. This corresponds to

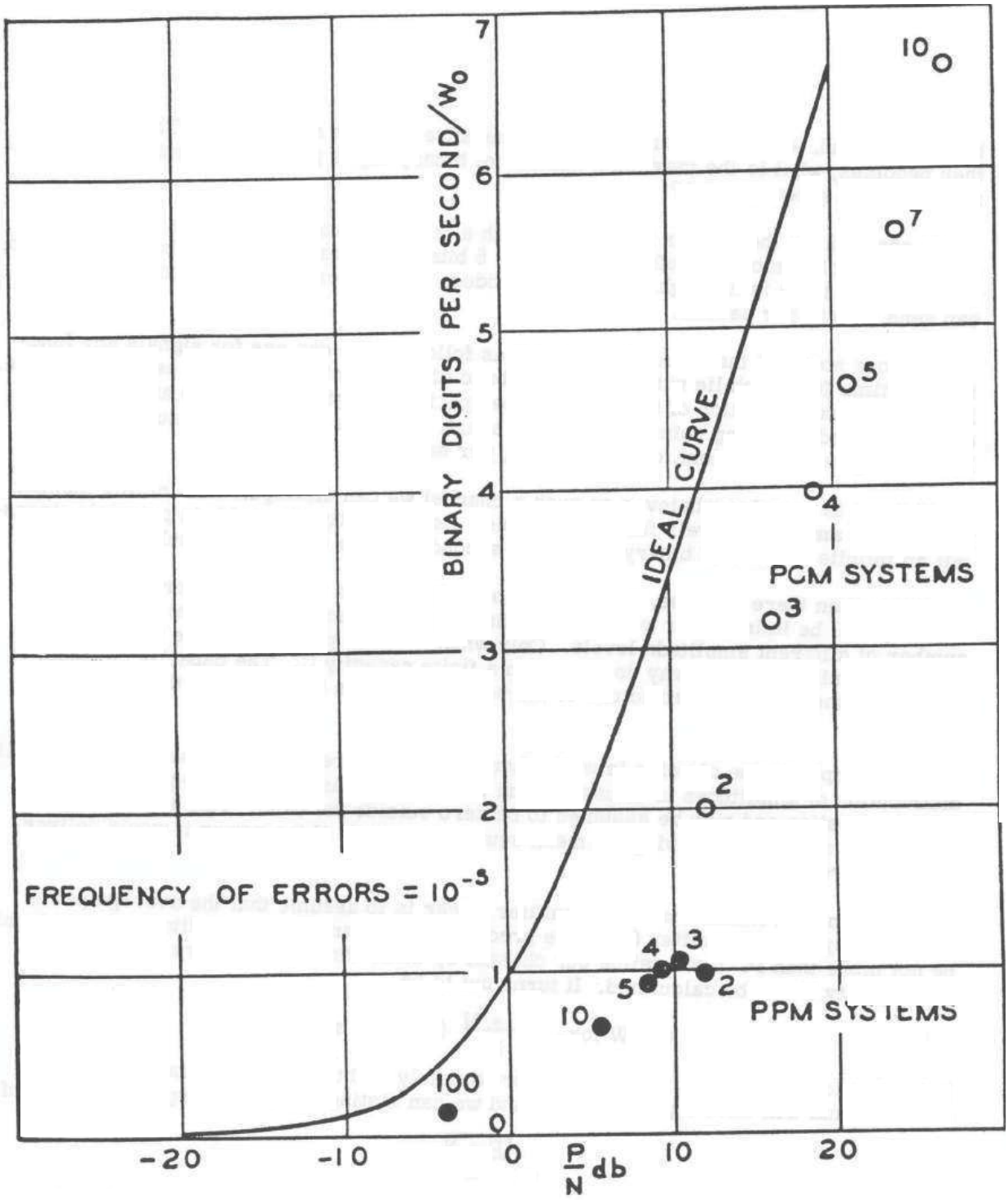
BINARY DIGITS PER SECOND/W_0
IDEAL CURVE
PCM SYSTEMS
PPM SYSTEMS
FREQUENCY OF ERRORS = 10^{-5}
$\frac{P}{N}$ db
-20
-10
0
10
20
0
1
2
3
4
5
6
7
100
10
7
5
4
3
2
10
5
4
3
2

Figure 4

$$\log_2 M = TW \log_2 \frac{P+N}{N}$$

binary digits in the time T or

$$C = W \log_2 \frac{P+N}{N}$$

binary digits per second. This formula has a much deeper and more precise significance than the above argument would indicate. In fact it can be shown that it is possible, by properly choosing our signal functions, to transmit $W \log_2 \frac{P+N}{N}$ binary digits per second with as small a frequency of errors as desired. It is not possible to transmit binary digits at any higher rate with an arbitrarily small frequency of errors. This means that the capacity is a sharply defined quantity in spite of the noise. These statements are proved by two different methods.[2]

The formula for C applies for all values of P/N. Even when P/N is very small, the average noise power being much greater than the average transmitter power, it is possible to transmit binary digits at the rate $W \log_2 \frac{P+N}{N}$ with as small a frequency of errors as desired. In this case $\log_2 (1 + \frac{P}{N})$ is approximated by $\frac{P}{N} \log_2 e = 1.443 \frac{P}{N}$ and we have approximately

$$C = 1.443 \frac{PW}{N}.$$

It should be emphasized that it is only possible to transmit at a rate C over a channel by properly encoding the information. In general, the rate C is only approached as a limit by using more and more complex encoding and longer and longer delays at both transmitter and receiver. In the white noise case the best encoding is such that the transmitted signals themselves have the structure of a white noise with power P. The difficulty with the approximate argument given for that case, and the reason it does not give a sharply defined capacity, is that the selection of signals is not optional. The distribution of amplitudes is not Gaussian as it should be.

4. Comparison of Ideal and Practical Systems.

In Figure 4 the curve is the function

$$\frac{C}{W} = \log (1 + \frac{P}{N})$$

plotted against P/N measured in db. It represents, therefore, the channel capacity per unit of band with white noise. The circle and points correspond to PCM and PPM systems used to send a sequence of binary digits and adjusted to give about one error in 10^4 binary digits. In the PCM case the number adjacent to a point represents the number of amplitude levels - 3 for example is a ternary PCM system. In all cases positive and negative amplitudes are used. The PPM systems are quantized with a discrete set of possible positions for the pulse, the spacing is $\frac{1}{2W}$, and the number adjacent to a point is the number of possible positions for a pulse.

The series of points follows a curve of the same shape as the ideal but displaced horizontally about 8 db. This means that with more involved encoding or modulation systems a gain of 8 db. in power could be achieved over the system indicated.

---

[2] See Shannon, C. E., "Mathematical Theory of Communication" and "Communication in the Presence of Noise."

Of course, as one attempts to approach the ideal, the transmitter and receiver required become more complicated and the delays increase. For these reasons there will be some point where an economic balance is established between the various factors. It may well be, however, that even at the present time more complex systems would be justified.

A curious fact illustrating the general misanthropic behaviour of Nature is that at both extremes of P/N (when we are well outside the practical range) the series of points in Figure 4 approach more closely the ideal curve. At very large P/N the PCM points approach to within $10 \log_{10} \frac{\pi e}{3} = 4.5$ db. of the ideal while with very small P/N the PPM points approach to within 3 db. The relation

$$C = W \log \left(1 + \frac{P}{N}\right)$$

can be regarded as an exchange relation between the parameters W and P/N. Keeping the channel capacity fixed we can decrease the bandwidth W provided we increase P/N sufficiently. Conversely, an increase in band allows a lower signal-to-noise ratio in the channel. The required P/N in db. is shown in Figure 5 as a function of the band W. It is assumed here that as we increase W, N increases proportionally:

$$N = W N_0$$

where N_0 is the noise power per cycle of band. It will be noticed that if P/N is large a reduction of band is very expensive in power. Halving the band roughly doubles the signal-to-noise ratio in db. that is required.

The channel capacity C can be calculated in many other cases. A general result that applies in any situation where the average transmitter power is limited to P is that the channel capacity is bounded by:

$$W \log \frac{P + N_1}{N_1} \leq C \leq W \log \frac{P + N}{N_1}$$

where N_1 is a parameter called the "entropy power" of the noise. It is defined as the power in a white noise having the same entropy as the actual noise. N is, as before, the average noise power.

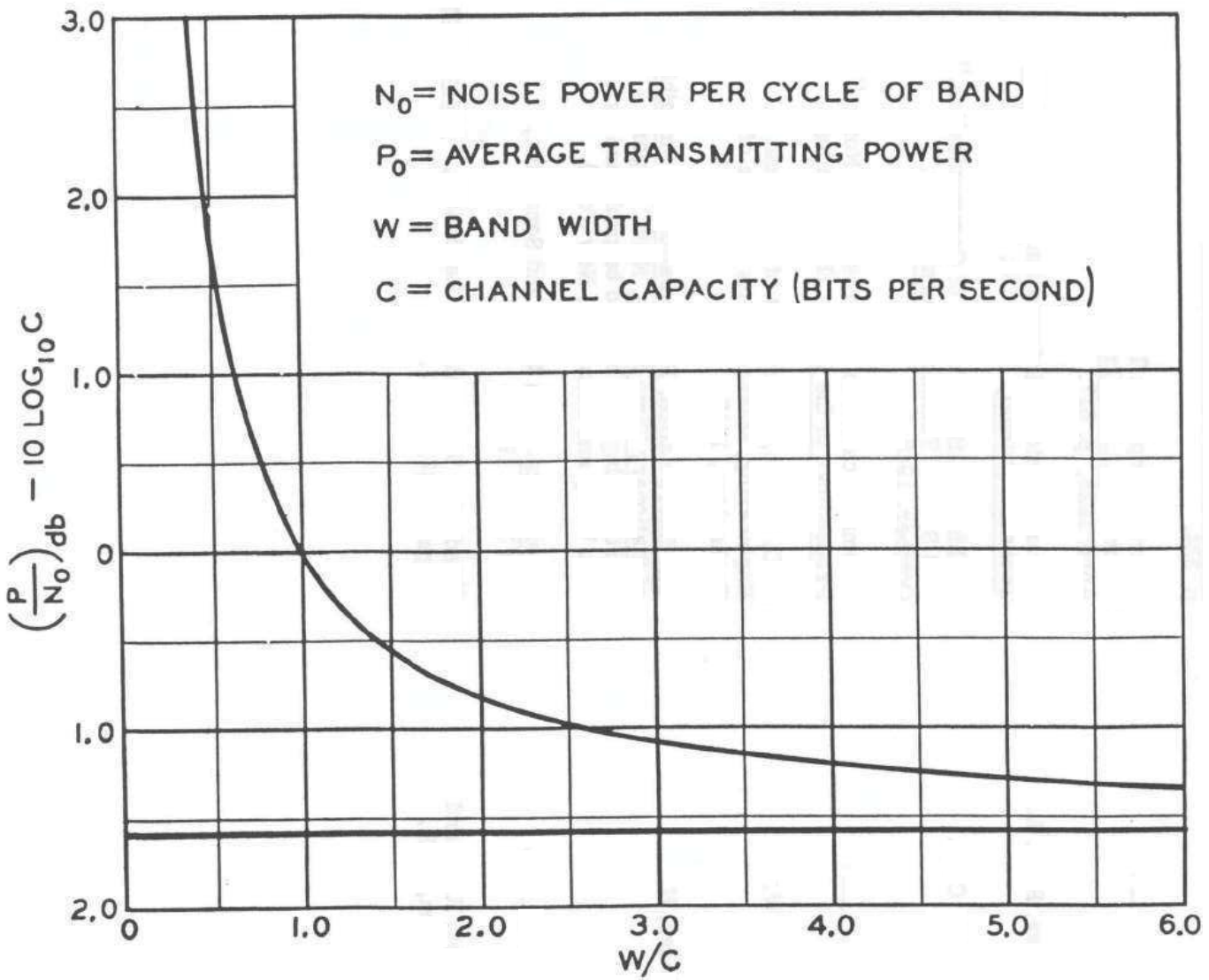
N_0 = NOISE POWER PER CYCLE OF BAND
P_0 = AVERAGE TRANSMITTING POWER
W = BAND WIDTH
C = CHANNEL CAPACITY (BITS PER SECOND)
$\left(\frac{P}{N_0}\right)_{db} - 10 \log_{10} C$
3.0
2.0
1.0
0
1.0
2.0
0
1.0
2.0
3.0
4.0
5.0
6.0
W/C

Figure 5

REFERENCES

Nyquist, H. "Certain Factors Affecting Telegraph Speed," Bell System Technical Journal, April 1924, p. 324.

———. "Certain Topics in Telegraph Transmission Theory," A.I.E.E. Transcripts, Vol. 47, April 1928, p. 617.

Hartley, R. V. L. "Transmission of Information," Bell System Technical Journal, July 1928, p. 535.

Shannon, C. E. "A Mathematical Theory of Communication," Bell System Technical Journal, July, October, 1948.

———. "Communication in the Presence of Noise," Proceedings of the I.R.E. (Forthcoming).

Tuller, W. G. Sc.D. Thesis, Department of Electrical Engineering, Massachusetts Institute of Technology, 1948.

Wiener, N. The Interpolation, Extrapolation and Smoothing of Stationary Time Series, NDRC Report (Forthcoming as a book to be published by John Wiley and Sons, Inc., New York).

———. Cybernetics. John Wiley and Sons, Inc., New York, 1948.

Bailey, R. D., and Singleton, H. E. "Reducing Transmission Bandwidth," Electronics, August 1948, p. 107.